AI-Detection Bias and False Positives: Comparing 2016 Human, 2026 AI, and 2007 Student Essays Across Common Detectors

10.6084/m9.figshare.31439995

Andrew Lehti
February 28, 2026

Part of the Collection of:
Cognitive Psychology and the Education System
https://doi.org/10.6084/m9.figshare.c.7532079

Physiological Advisory: The Biology of Belief
The following text challenges established paradigms and may contradict your first learned beliefs. This encounter will likely trigger a biological response. Because modern education conditions the mind to associate being wrong with punishment, your brain perceives conflicting data as a survival threat. This activates **Cognitive Impasse**, a physiological defense mechanism where cognitive inertia, dissonance, bias, and avoidance function as a single, integrated framework. This impasse often begins with an automatic urge to ridicule, dismiss, and laugh at new ideas or information that contend with your established beliefs (academic, cultural, professional, religious) to relieve internal tension, a reflex that frequently transitions into anger; but self-awareness can mitigate and ease these symptoms.

If you persist, this biological defense may escalate into visceral symptoms like a tight gut, headaches, or profound feelings of guilt and impending doom. Before this peak, a stage called **Cognitive Dissonance Avoidance** occurs, manifesting as a sudden compulsion to escape. You may feel abruptly sleepy, confused, or desperate for distraction. This is not boredom, but rather a survival mechanism masking itself as fatigue to protect your mind from the visceral pain of *possibly being wrong*. During this stage, the brain can become temporarily impaired that you may lose the ability to comprehend the evidence or even the literal meaning of the words being read or said. To understand this text, you must recognize that a feeling of exhaustion or doom is not evidence of falsehood, nor is it evidence of truth; but rather the feeling of a mental boundary developing.

*Genuine insight is rarely a matter of raw brilliance, but rather the grit to **"stay with the questions much longer."** For many, **"the only thing that interferes with my learning is my education,"** as formal schooling often prioritizes obedience over the true **"measure of intelligence,"** which is **"the ability to change your mind,"** as reflected in the beliefs of Albert Einstein. This mental flexibility is accessible to anyone, whether through formal study or as an autodidact, yet it remains rare within the rigid dogma of academia or any corpora. As Max Planck famously observed, **"science progresses one funeral at a time,"** because the hardest part of learning is often the willingness to outlive the paradigms we were taught.*

Lehti, Andrew (2025). Cognitive Impasse [Visual]: Stages of Mental Rigidity, Dissonance, and Bias. figshare. Figure.
https://doi.org/10.6084/m9.figshare.30016195

Updated 2/28/2026

Keywords: AI detection reliability, AI detection false positives, AI detection bias, AI detector accuracy, AI authorship detection, AI detection tools comparison, AI detection in academia, human vs AI writing comparison, AI plagiarism detection accuracy, inter-detector variability, AI detection statistical analysis, AI detection ethics, false academic misconduct claims, AI detection validation study, formatting bias in AI detection, AI detection inconsistency, AI detection reproducibility, AI-assisted writing vs AI authorship, can AI detectors be trusted, are AI detectors reliable

Methodology

This research investigates motives, behaviors, and observations by deliberately delaying exposure to academic and institutional findings. By utilizing an extrapolative trial by error approach, the study minimizes the anchoring biases and standardized obedience that typically restrict the thought process. This isolation ensures that the initial synthesis of data remains unpolluted by the biological defense mechanisms, such as the ***Cognitive Impasse***, often triggered when a researcher is tethered to a specific institutional paradigm. Following the independent completion of the study, academic material is reviewed comprehensively to evaluate alignment. Typically, the findings correspond closely with established principles, with exceptions limited to nuanced insights not yet formally recognized but widely observed in society

Lehti, Andrew (2024). Extrapolative Trial by Error. Journal contribution.
https://doi.org/10.6084/m9.figshare.27643080

Google NotebookLM; A Guide to Exploring Complex Works

These works are dense and interconnected, revealing deeper structures only through careful study. To navigate them, readers are encouraged to use **Google NotebookLM**, a free AI tool that synthesizes information across multiple texts. It highlights connections, answers questions with accuracy, and even creates interactive AI-driven podcasts, allowing users to explore ideas dynamically and with greater clarity.

notebooklm.google.com

About Author

Andrew Lehti is an independent researcher and polymath autodidact who has dedicated over 35,000 hours to **Metopedia**, a research journey designed to diagnose the **Cognitive Impasse**. This research identifies the impasse as a physiological defense cycle that forces the mind to reject information contradicting first learned beliefs. The work argues that modern education is a system of **Standardized Obedience** that conditions the brain to perceive error as a survival threat. By associating intellectual deviation with punishment, this systemic conditioning triggers biological resistance such as the involuntary urge to mock new ideas or the onset of brain fog to avoid the pain of cognitive dissonance.

Lehti utilizes a methodology termed **Extrapolative Trial by Error**, which intentionally bypasses academic literature during initial investigations to prevent anchoring bias. This approach has led to the development of the **Canonical Order of Operations** in mathematics and the use of the **Codex Amiatinus** to identify critical linguistic mistranslations in historical religious texts. He has authored proprietary tools including **Image Degradation Analysis (IDA)** for forensic inquiry and the **Polyhedral Index Partition (PIP)** for calculating integer partitions. Through these diverse applications, the work seeks to provide the cognitive tools necessary to dismantle the collective mental rigidity that governs both individual behavior and institutional dogma.

The materials published by Andrew Lehti reflect an interdisciplinary approach spanning mathematics (classical, algebraic, and computational), investigative history, linguistics with an emphasis on translation and reconstruction of religious texts, cognitive psychology and philosophy of mind, and the philosophy of education, objectivity, exploration, and metacognition.

Selected Works:

1. Lehti, Andrew (2024). Selective-Mindedness: An Introduction and the Illusion of Open-Mindedness. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.27642519

2. Lehti, Andrew (2026). Ancient Latin Translation Fixes and the Reconstructed Relationship Between Rome and Scripture. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.31272391

3. Lehti, Andrew (2024). When Death or Loss Makes Us Laugh: Unraveling the Emotional Paradox and Exploring the Connection Between Grief and Humor. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.28014581

4. Lehti, Andrew (2024). The Reptilian People in Authority: Basilicas, Basilisks, and an Allegory. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.28016237

5. Lehti, Andrew (2026). Image Degradation Analysis (IDA): Forensic Applications for Light Physics, White Paper. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.31323694

6. Lehti, Andrew (2025). An Exploratory of Universal Cosmic Descent. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.28454402

7. Lehti, Andrew (2024). Cognitive Impasse and the Puppet Master of Society: A Framework of Mental Rigidity. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.28014626

10.6084/m9.figshare.31439995

Abstract

The integration of generative language models into academic and professional settings has driven adoption of AI-detection software. Universities use these tools to assess authorship, but their reliability remains contested. This study evaluates several common detectors by comparing their assessments of three texts: a 7,062-word polished human essay (2016), a 1,100-word AI-generated essay (2026, mixed formal–informal), and a 900-word seventh-grade human essay (2007). Results show extreme variability across detectors, substantial false positives on human writing, and high inter-tool disagreement. Notably, the 2016 human essay scored higher for AI than the AI-generated essay. These findings raise serious concerns about institutional reliance on AI detectors for academic judgment.

> *The average AI-detected percentage for this abstract was 48.94%; the two lowest were 3.65% and 16.86%, while the two highest were 100% and 100%.*

10.6084/m9.figshare.31439995

Introduction

Educational institutions have been trying to find ways to detect AI-generated content as large language models increase. AI detection systems assert their capability to use characteristics such as "perplexity," "syntactic regularity," "lexical predictability," and "structural coherence" as a means of distinguishing human-generated content from machine-generated content.

Text produced algorithmically can resemble formal writing in academic content because of characteristics such as tone, lack of slang, technical vocabulary, use of semicolons and em-dashes, structured text arrangement, and punctuation consistency. The challenge is that these systems are trained on large amounts of formal and informal writing merely mimicking academic and semi-formal writing styles.

This paper seeks to evaluate the degree of accuracy at which public and commercial AI detection systems can distinguish AI-generated content from older human-generated content.

Essays and Method

Three texts were evaluated:

1. **2016 Human Essay**
    - 46,000 characters (7,062 words)
    - Polished, semi-academic style
    - Formal narrative beginning: “People in science, especially in academia, often become overly critical of trivial details…”
2. **2026 AI Essay**
    - 1,115 words
    - Mixed formal and informal register
    - Semi-formal narrative beginning: “Sensory input is transduced into neural representations, structured by prior expectations…”
3. **2007 Human Essay (Seventh Grade)**
    - 898 words
    - Informal, beginner
    - Informal narrative beginning: “I want to tell you about a time in my life that…”

Each document was submitted independently to multiple AI-detection systems. Reported percentages reflect each tool’s claim of AI-likelihood. Where tools required chunking, rounded averages were calculated.

Update: See Appendix C. for additional detector scores and revised insights

Added: 2026 Human Essay, 2026 Human Comment

Results

Large Sample Comparison (2016 Human Essay vs. 2026 AI Essay)

Detector	2016 Human	2026 AI
Copyleaks	96.8%	99.9%
ZeroGPT	54.51%	31.67%
GPTZero	98%	92%
Gemini 3 Thinking	80–90%	90–95%
ChatGPT Extended Thinking	35–45%	75–85%
Quillbot	29%	54%
Sapling AI	100%	29%
Grammarly	45%	15%
AIDetector	2.35%	4.75%
Scribbr	37.5%	26%
undetectableAI	67%	72%
originalityAI	100%	100%
Pangram	100%	100%
NoteGPT	28.9%	53.26%
GPTInf	100%	100%
eduwriter AI	49%	29%
Winston AI	99%	61%

Note: *Quilbot serves as the backend for Scribbr; however, because of a non-subscription status, the word limit was restricted, resulting in differences in scores.*

*See **Appendix A.** and **Appendix B.** for additional results*

2016 Human Essay Statistics

- Mean: 66.59%
- Median: 67.00%
- Standard Deviation: 32.78%
- Range: 2.35% to 100%

2026 AI Essay Statistics

- Mean: 59.25%
- Median: 57.50%
- Standard Deviation: 33.69%
- Range: 4.75% to 100%

Observations

1. The human essay was scored higher on average than the AI-generated essay.
2. Both samples produced ranges spanning from near-zero to 100% AI classification depending on detector.
3. Standard deviations were extremely high in both cases, indicating poor consistency across tools.
4. Several platforms appear structurally aligned with paraphrasing or rewriting services, raising potential conflict-of-interest concerns.

Informal Writing Control: 2007 Essay

The seventh-grade essay produced markedly different outcomes. Most detectors classified it as overwhelmingly human. While only these differences were observed:

- AIDetector: 1.75%
- Scribbr: 1%
- Quillbot: 1%
- Sapling: 26.4%
- NoteGPT: 16.15%
- GPTZero: 2%
- WinstonAI: 5%
- undetectableAI: 9%
- ZeroGPT: 16.15%
- Eduwriter AI: 16%
- OpenL IO: 16%
- YouScan: 15%

When evaluated directly by major models:

- ChatGPT Extended Thinking: 10–25% AI (cited messy flow, irregular punctuation, inconsistent structure)
- Gemini 3 Pro Thinking: 5–10% AI (cited poor organization, punctuation errors, cyclical structure)

This suggests detectors may interpret imperfection and structural irregularity as human markers.

Update: See Appendix C. for additional detector scores and revised insights

Stylistic Ecosystem Convergence

Beyond numerical inconsistency, a broader structural issue emerges; generative AI systems are trained on the very discourse ecosystems they are later asked to emulate or avoid. **Contemporary large language models are trained on vast corpora of online and academic writing, including:**

- Reddit essays and comment threads
- Reddit answers and long-form discussions
- Quora responses
- Academic papers and discussions
- Professional blog writing
- Structured forum articles

As a result, GPT output often resembles the arithmetical center of present-day internet-native writing. AI detectors are therefore tasked with distinguishing between **(a)** human writing shaped by the digital ecosystem and **(b)** machine writing trained on that same ecosystem.

This convergence substantially weakens discrimination boundaries. If modern human discourse is structurally similar to the averaged linguistic signal of training corpora, then classification tools may detect "internet-era coherence" rather than artificial authorship.

Formatting and Structural Bias

Additional testing revealed that formatting variables alone can meaningfully alter AI detection scores. **Observed pattern shifts conclude that greater length mixed with:**

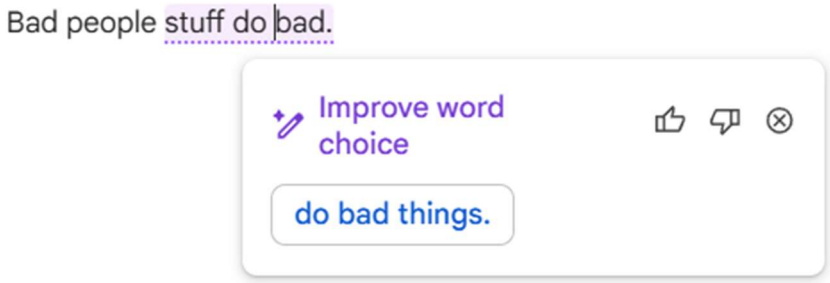
- Increased use of em dashes correlated with some elevated AI percentages, especially in simple systems.
- Frequent semicolons and colons increased AI attribution likelihood in some systems.
- Structured section headings and markdown formatting raised AI flags in some systems.
- Higher vocabulary density and lexical precision increased detection probability.
- Consistent punctuation and grammatical confidence produced stronger AI classification signals.

In repeated tests, reformatted versions of identical content—without semantic alteration—produced higher AI scores solely due to structural cleanliness. Markdown layouts were particularly disfavored by several detectors, suggesting that formatting heuristics may influence scoring. This posits an important practical concern: detectors may be responding to visual and structural regularity rather than generative origin.

AI Authorship vs. AI-Assisted Revision

A significant distinction frequently omitted in AI-detection debates is the difference between full generative authorship and revision-based assistance. **These are not equivalent practices**, yet most detection systems treat them as interchangeable.

Full AI authorship involves prompting a model to generate an entire piece of content work from minimal input. It does not involve structured investigation, hypotheses, nor evidence gathering; In contrast, AI-assisted revision involves a human author composing the substance of a text from their data and hypotheses, and then using software tools to improve clarity, flow, or grammatical consistency.



In the modern writing space, there are many embedded AI revision tools that assist the writer. For instance, there are many word processing tools, online document editors, and especially email clients that can clarify the text with one click of the mouse, making slight changes to the sentence structure while keeping the meaning of the sentence the same. **They have the following characteristics:**

- Smooths out awkward phrasing
- Removes redundancy
- Standardizes punctuation
- Improves syntactic coherence
- Modifies the vocabulary for clarity

These micro-adjustments introduce measurable stylistic regularization. When applied across a document, even lightly, they can increase statistical uniformity—precisely the feature many AI detectors interpret as algorithmic generation.

This creates a convergence mechanism:

- Human-authored draft
- AI clarity smoothing
- Increased structural regularity
- Elevated AI detection probability

Such revision tools are now normative rather than exceptional. Many writers routinely use them for **tone adjustment**, **professionalism**, **efficiency**, or **readability**. If detection systems do not differentiate between authorship-level generation and clarity-level revision, then ordinary writing practice becomes algorithmically suspect.

Importantly, revision does not displace authorship. The argument, ideas, structure, and intellectual content remain human-derived. Clarity-adjustments function more like editorial polishing than generative composition.

However, because of their high level of consistency in syntactical distribution, it's possible that systems of revisions might produce texts that are statistically similar to those generated by large language models. This might cause detection systems to identify stylistic similarity instead of origin.

In such a world, it becomes difficult to distinguish between assistance and authorship. If AI tools are incorporated into popular platforms that can easily correct sentence structure, punctuation rhythm, and vocabulary use, then it becomes difficult to distinguish between machine and human authorship—no matter how human the intellectual content might be.

In such a world, it becomes difficult to define a system of classification.

The Theoretical Impossibility of Stable AI Detection

Beyond empirical inconsistency lies a deeper structural problem: the difficulty of constructing a stable AI-detection algorithm in a convergent linguistic ecosystem. The challenge is not merely technological, but recursive one.

1. Human-AI Style Convergence

Humans learn languages from other humans. Increasingly, humans also learn from AI-generated text—directly through interaction, indirectly through exposure. As AI-generated prose permeates online forums, academic discussions, professional communication, and social discourse, **individuals read and internalize these patterns.**

Language-acquiring operates partly through mimicry and implicit statistical absorption. If AI models produce syntactic regularity, structured argumentation, and predictable punctuation patterns, and humans repeatedly encounter those patterns, **humans will subconsciously incorporate them into their own writing.**

This produces a feedback loop:

1. AI produces structured text
2. Humans read and absorb structural features
3. Humans adopt those features
4. Human writing increasingly resembles AI training distributions

As convergence increases, separability decreases. An algorithm attempting to isolate a uniquely "machine" signature must contend with a population whose stylistic patterns are co-evolving with the machine.

If the distribution overlap approaches equilibrium, discriminative classification becomes mathematically unstable. It becomes a game of cat and mouse that never ends and will never have non-error results.

2. Pre-Existing Internet Convergence

Even before modern generative models, linguistic convergence was accelerating. **With the rise of the internet**, millions of writers were exposed to shared stylistic norms: forum layouts, blogging conventions, standardized academic summaries, email etiquette, technical explanations, list formatting, markdown, and citation styles.

Writers learned from each other at unprecedented scale. The global network compressed linguistic variability and amplified certain norms.

Large language models were later trained on this internet-era corpus. **As a result:**

- AI models are trained on human-written text shaped by digital convergence.
- Humans now write in styles partially shaped by digital exposure.

Detecting AI becomes an attempt to distinguish a statistical average of human language from individual human instances of that same average.

3. The Collapse of Stable Linguistic Markers

Early detection methods relied on heuristics such as:

- Low perplexity
- Reduced entropy
- Structural symmetry
- Predictable lexical transitions

However, as models improve, they introduce controlled randomness. Simultaneously, human writers increasingly rely on clarity tools, grammar correction systems, predictive typing, and minor AI revisions.

This leads to two simultaneous effects:

- AI systems increase variability to mimic human irregularity.
- Human writing becomes more regular through software assistance.

The distributions move toward each other from opposite directions.

If AI becomes more human-like while humans become more algorithmically polished, the decision boundary between classes erodes. Classification systems depend on measurable separation; convergence reduces measurable separation.

4. Recursive Training Dynamics

A further complication arises from recursive data ingestion. Contemporary language models are increasingly trained on corpora that may include AI-generated text. Meanwhile, human writers exposed to AI output produce derivative patterns, which feed back into future datasets.

The training corpus becomes a blend of:

- Original human text
- Human text influenced by AI
- AI-generated text influenced by human text

Once this entanglement reaches sufficient density, identifying origin through surface-level statistical analysis becomes intractable. The signal of authorship origin becomes indistinguishable from stylistic influence.

5. Epistemic and Computational Limits

From a classification theory perspective, **detection requires:**

1. A stable signature unique to AI output.
2. A distribution distinct from human-authored text.
3. Low overlap between the two populations.

In a converging ecosystem, all three conditions fail.

If stylistic convergence increases over time, any fixed algorithm will degrade in accuracy. If adaptive retraining occurs, retraining will rely on evolving datasets that already contain blended authorship signals.

The system chases a moving boundary that may ultimately disappear.

Theoretical Argument

If human language evolves through exposure, imitation, and technology, and we, as humans, have exposed AI to human language, which in turn has impacted human language, we have a co-adaptive relationship.

With the assumption of sustained convergence, the idea of a definitive, future-proof algorithm to detect AI becomes a conceptual paradox.

The question becomes less about detection, as in "Can we detect AI?" It becomes more of a question: "Can we detect the source of a linguistically entangled ecosystem?" Ultimately, as we continue to converge, the distinction becomes less relevant to computation or philosophy itself.

Detector Bias and the "Polish Penalty" Hypothesis

Across comparative trials, a consistent directional pattern appeared:

- Informal, inconsistent punctuation, and structurally inconsistent writing scored overwhelmingly human.
- Clean, structured, academically formatted writing scored disproportionately AI.

The 2007 seventh-grade narrative, which contained irregular sentence flow, inconsistent punctuation, repetition, and structural flaws received near-universal "human" classification.

By contrast, a polished 2016 academic essay written years before generative AI became mainstream received a mean AI detection score of 66.59%, with some detectors assigning 100%.

This suggests what may be termed a "polish penalty" effect:

- As structural competence increases, AI attribution probability increases.

If detection systems implicitly equate regularity, greater vocabulary competence, coherence, and punctuation consistency with algorithmic generation, then skilled human writers may be systematically disadvantaged.

Humanization and the Paradoxical Score Shift

In additional controlled tests, a personally written academic text—verified human-authored—initially received a mean AI score of 44.7%.

The same text was then processed through multiple AI "humanization" or "AI bypass" tools designed to reduce detector sensitivity. **The result was as expected:**

- The average AI score **increased to 76.4%** on the humanized text

In other words, the attempt to make clearly human text appear "more human" **made it appear more AI-generated.**

This reversal suggests instability in detection heuristics. If stylistic deviation intended to lower AI scores instead raises them, the internal detection signals may be highly sensitive to surface-level statistical changes rather than true authorship markers.

Bias Implications

Taken together, the ecosystem convergence, formatting sensitivity, vocabulary effects, punctuation weighting, and humanization paradox indicate that:

- AI detectors do not operate on a stable, widely agreed linguistic signature of artificiality.
- Structural confidence may be penalized.
- Formatting and presentation style can alter classification independent of authorship.
- Detector outputs vary not only across tools but across minor formatting adjustments within the same text.

If formal academic writing, markdown formatting, em-dash usage, advanced vocabulary, and punctuation proficiency systematically increase AI scores, then detector bias becomes entangled with writing competence.

Discussion

The findings reveal three systemic issues:

1. Stylistic Convergence

AI models have been trained on copious amounts of internet content as well as academic literature. As a result, both well-constructed academic content and well-constructed content in general, whether human- or machine-generated shares predictable linguistic patterns. The detection tools, it seems, use formal coherence as a probabilistic criterion of artificiality.

2. Extreme Variability

Inter-detector variability borders on methodological chaos. If the range of AI detection lies between 2% and 100%, the reliability of such detection tools becomes statistically dubious. The high standard deviation of 30% or more in both groups indicates a lack of classifier agreement.

3. Incentive Structures

Many of the AI detection tools offer rewriting or "AI bypass" services, which are prominently featured on their websites. This conflict of interest, whether real or perceived, has implications for the overall integrity of institutional and personal trust.

Implications for Academia

If a polished 2016 human essay—written years before widespread generative AI adoption—can receive a higher average AI score than contemporary AI output, then detector validity is not just unreliable, but unethical.

Universities and teachers using such tools risk:

- False accusations of misconduct
- Administrative disputes grounded in probabilistic models
- Penalization of students who write in formal academic tone

Because formal writing includes structured argumentation, controlled vocabulary, disciplined punctuation, and predictable layout, detectors may conflate academic competence with artificial generation.

Conclusion (Human)

The findings of this comparative analysis show that the existing AI detection tools do not possess the required level of reliability for serious academic enforcement. The extreme variability, high false-positive rates, and detectable stylistic bias all contribute to a lack of confidence in such AI detection tools for academic institutions.

This is not a technical problem; it is an epistemic problem. If we ask a machine learning system trained on formal writing to detect formal human writing versus formal machine writing, it is no longer a meaningful distinction; it is a guessing game.

Until such time as transparent validation studies, standardized benchmarks, and false-positive audits are conducted, AI detection should not be used as determinative evidence.

44% AI Detected Average

Conclusion (Humanized)

This analysis shows current AI-detection systems lack reliability for high-stakes academic enforcement. Variability, high false-positive rates, and stylistic bias undermine institutional confidence. The problem is not merely technical but epistemic: when models trained on formal language try to distinguish formal human and machine writing, the task reduces to probabilistic guesswork. Without transparent validation, standardized benchmarks, and false-positive audits, AI detection should not serve as decisive evidence in educational or professional decisions.

76% AI Detected Average

Revised Conclusion

Appendix C shows that structured writing, whether semi-formal or even informal with proper grammar and structure, produces misclassification rates that approach parity. A little less than 50% of clearly structured human-written text is identified as AI, whereas almost the same percentage of AI-generated material is identified as human. In effect, once structure is established, detection accuracy approaches random performance. The probability of misclassification becomes statistically equivalent to a coin flip.

Appendix A. Qualitative Control Case: Reddit Comment Classified as Human

A.1 Text Sample ("Clearly Not AI")

"I have been accused on reddit of bing Ai. I don't use Ai. I have no need for Ai. I am an artist who can draw, sketch, paint, sculpt, airbrush, etc. I write so much but I am in no way a writer in the sense of professional or as a hobby. I studied to use illustrator, photoshop, 3d modeling and animation in Maya, etc. I have never wanted to rely on others or be a slave to my incompetence. I do not use my diploma. I drive semi as a CDL construction contractor. I am self employed. There is no satisfaction in having anything done for me. I do all my maintenance and repairs. If I could I would raise my own food. I would build my own home. I want to do all of it. None of it requires Ai. I have asked Ai things and I learned quickly that it knew nothing. Back when I was in college I made a license plate design which I then provided to another graphic designer who was looking for inspiration. I came up with in less than a minute. I expected this person to change it up. They submitted it as is and is now one of the main license plates in my state. You can't get such satisfaction from using Ai. I see that plate daily."

A.2 Detector Outputs

Detector	Result
Copyleaks	0% AI
ZeroGPT	**3.24% AI**
GPTZero	0% AI
Quillbot	0% AI
Detecting-AI	**29.1%**
Sapling AI	0% AI
Grammarly	0% AI
AIDetector	**2.37% AI Detected**
Scribbr	0% AI
undetectableAI	**5% AI**
originalityAI	0% AI
Pangram	0% AI
NoteGPT	**3.27%**
GPTInf	0%
eduwriter AI	**3% AI**
Winston AI	**2%**
OpenL IO	**3% AI**
YouScan	**15% AI**

A.3 Model-Based Rationale Summaries (As Reported)

Gemini 3 Thinking: 0% AI

> Rationale: idiosyncratic, highly specific narrative, human errors, non-linear structure, explicit critique framed as "nothingness."

ChatGPT Extended Thinking: 0–15% AI

> Rationale: typos and slips, rant-like cadence, repetition, specific personal details, uneven phrasing.

A.4 Interpretation (Control Value)

This sample functions as a qualitative control in which most detectors converge on "human" classification. The text contains multiple features commonly treated as human markers: non-linear narration, uneven pacing, repetitive emphasis, inconsistent capitalization, minor errors, and dense personal specificity. Notably, a minority of tools still assigned non-trivial AI likelihood (e.g., Detecting-AI at 29.1%, YouScan at 15%), reinforcing the broader finding that even strongly human-coded text can yield disagreement across detectors.

Appendix B. Qualitative Control Case: 2026 Human Essay

First-person, informal/semi-formal, unpublished essay, 3,502 words, 22,372 characters; no tools, polish, correction, or revision; first 275 words:

B.1 Text Sample

"Reddit is indeed a hivemind, or rather, the same people hang out in the same areas you are likely hanging out in due to shared interests, so there is always overlap with the same users that varies from subreddit to subreddit. When I first started reviewing -5^2, it was due to a result on Facebook around the beginning of COVID. Users attacked me when I claimed that the result of -5^2 was -25.

Mathematics, a field in which I do well, was being mocked with incorrect answers and terrible optimizations/conventions taught by ineffective teachers, and practically everyone agreed on the incorrect convention in that discussion. When I went to Reddit, it took a couple posts for users to start commenting the same correct answers one or two users wrote before; but after the third or so posting, people began to recognize me: "Oh, it's you." Just like in the Facebook discussion, the same rigid responses were delivered again.

Between my initial Reddit post and the Facebook episode, I conducted extensive research to see whether I was truly in error. The sheer amount of people who disagreed with me made me feel as though I had done something incorrectly. At the time, I didn't think the collective could be so wrong.

Nevertheless, it was an excellent learning experience because it helped me develop the Canonical Order of Operations, which I only published in 2024, and finalized in 2026 following a few years of extensive historical and mathematical research.

> *Lehti, Andrew (2024). The Canonical Order of Operations: A Separate Framework. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.27661734.v3*

Here, the publishing date is important, but not because it concludes the narrative…"

B.2 Detector Outputs

Detector	Result
Copyleaks	77.4% AI
ZeroGPT	7.1% AI
GPTZero	38.31% AI
Quillbot	13% AI
Detecting-AI	30.6%
Sapling AI	27.6% AI
Grammarly	30% AI
AIDetector	12.47% (80%+ AI probability)
Scribbr	19%
undetectableAI	77% AI
originalityAI	100% AI
Pangram	39% AI
NoteGPT	7.87%
GPTInf	11% AI (100% mixed)
eduwriter AI	17% AI
Winston AI	3%
OpenL IO	7% AI
YouScan	85% AI
Detector IO	21% AI
Content Detector AI	18% AI
Decopy AI	57% AI
Dechecker	8% AI

Note: Quilbot serves as the backend for Scribbr; however, because of a non-subscription status, the word limit was restricted, resulting in differences in scores.

2016 Human Essay Statistics

- Mean: 33.60%
- Median: 24.3%
- Standard Deviation: 28.31%
- Range: 3% to 100%

B.3 Model-Based Rationale Summaries (As Reported)

Gemini 3 Thinking: 30-40% AI

Rationale: sustained academic tone, repeated claim–implication–restatement cycles, high structural symmetry, argumentative momentum

ChatGPT Extended Thinking: 65% AI

Rationale: embedded personal chronology, platform-specific anecdotes, coined terminology, variable sentence rhythm, occasional informal compression, narrative irregularities

B.4 Interpretation (Control Value)

Residual AI scores assigned by a subset of tools illustrate the instability of detector thresholds. Even when multiple human-coded markers are present, probabilistic models may still register patterned coherence or formal synthesis as generative. The disagreement across systems therefore reinforces the larger conclusion.

Appendix C. Full Detector Data Tables

Appendix C shows the inclusion of extra detector outputs, which turn many single-condition data into complete five-condition profiles. When partial rows are substituted with scored data instead of absent values, the structural result is improved comparison across tools and a quantifiable increase in variance. Variance rises in the Human Comment condition, exposing disagreement that was previously obscured by missing entries, but global central tendency only slightly changes.

C.1 All Detector Outputs

Detector	2026 Human Essay	2007 Human Essay	2016 Human Essay	2026 AI Essay	2026 Human Comment
AIDetector	12.5%	1.8%	2.4%	4.8%	2.4%
ChatGPT	65.0%	17.5%	40.0%	80.0%	7.5%
Content Detector AI	18.0%	0.0%	73.0%	0.0%	23.0%
Copyleaks	77.4%	0.0%	96.8%	99.9%	0.0%
Dechecker	8.0%	16.0%	16.0%	31.0%	3.2%
Decopy AI	57.0%	46.0%	51.0%	34.0%	32.0%
Detecting-AI	30.6%	42.6%	48.3%	67.9%	29.1%
Detector IO	21.0%	0.0%	21.0%	70.0%	0.0%
GPTInf	11.0%	0.0%	100.0%	100.0%	0.0%
GPTZero	38.3%	2.0%	98.0%	92.0%	0.0%
Gemini	35.0%	7.5%	85.0%	92.5%	0.0%
Grammarly	30.0%	1.0%	45.0%	15.0%	0.0%
NoteGPT	7.9%	16.2%	28.9%	53.3%	3.3%
OpenL IO	7.0%	16.0%	37.0%	42.0%	3.0%
Pangram	39.0%	0.0%	100.0%	100.0%	0.0%
Quillbot	13.0%	1.0%	29.0%	54.0%	0.0%
Sapling AI	27.6%	26.4%	100.0%	29.0%	0.0%
Scribbr	19.0%	1.0%	37.5%	26.0%	0.0%
Winston AI	3.0%	5.0%	99.0%	61.0%	2.0%
YouScan	85.0%	15.0%	15.0%	25.0%	15.0%
ZeroGPT	7.1%	16.2%	54.5%	31.7%	3.2%
eduwriter AI	17.0%	16.0%	49.0%	29.0%	3.0%
originalityAI	100.0%	0.0%	100.0%	100.0%	0.0%
undetectableAI	77.0%	9.0%	67.0%	72.0%	5.0%
Mean	**33.60%**	**3.67%**	**66.59%**	**61.18%**	**5.49%**
Median	**24.30%**	**1.00%**	**67.00%**	**61.00%**	**2.19%**
Std. Dev.	**28.31%**	**7.00%**	**32.78%**	**33.57%**	**9.42%**

Context: *2007* ***Seventh Grade*** *Human Essay* and *2026 Human* ***Reddit*** *Comment:* poor structure, inconsistent, poor flow. *2026 Human Essay*: first-person, informal/semi-formal, experience driven. *2016 Human Essay:* academic semi-formal, third-person. *2026 AI Essay:* semi-formal, fully generated AI.

C.2 Detector Means

Detector	Mean by Row	Human Error Rate	AI Error Rate**
AIDetector	4.70%	7.89%	95.25%
ChatGPT	42.00%	48.33%	20.00% **
Content Detector AI	22.80%	53.33%	100.00%
Copyleaks	54.80%	58.07%	0.10% **
Dechecker	14.80%	16.52%	69.00%
Decopy AI	44.00%	83.33%	66.00%
Detecting-AI	43.70%	69.60%	32.10%
Detector IO	22.40%	14.00%	30.00%
GPTInf	42.20%	37.00%	0.00% **
GPTZero	46.10%	46.10%	8.00% **
Gemini	44.00%	42.50%	7.50% **
Grammarly	18.20%	25.33%	85.00%
NoteGPT	21.90%	20.91%	46.74%
OpenL IO	21.00%	23.00%	58.00%
Pangram	47.80%	46.33%	0.00% **
Quillbot	19.40%	14.33%	46.00%
Sapling AI	36.60%	51.33%	71.00%
Scribbr	16.70%	19.17%	74.00%
Winston AI	34.00%	37.67%	39.00%
YouScan	31.00%	53.33%	75.00%
ZeroGPT	22.50%	29.16%	68.33%
eduwriter AI	22.80%	30.33%	71.00%
originalityAI	60.00%	66.67%	0.00% **
undetectableAI	46.00%	56.00%	28.00%
Mean	**34.10%**	**40.11%**	**38.82%**

***The AI Error Rate is determined as "how much probability" the detector failed to assign to the AI class; nevertheless, with only one sample size for each, this should be considered insufficient when considering only the detector.*

C.3 Detector Error Rates by Category

Human Detection Error Rate	AI Detection Error Rate	Semi-Formal Error Rate	Low-Quality Writing Error Rate
26.95%	45.4%	45.8%	8.1%

Low-Quality Writing includes the 2007 Human Essay and the 2026 Human Comment. Semi-Formal includes the 2026 and 2016 Human Essays. Human Detection Error Rate includes all human-produced content.

Appendix D. Detector Graphs

D.1 Detector AI-Probability Heatmap

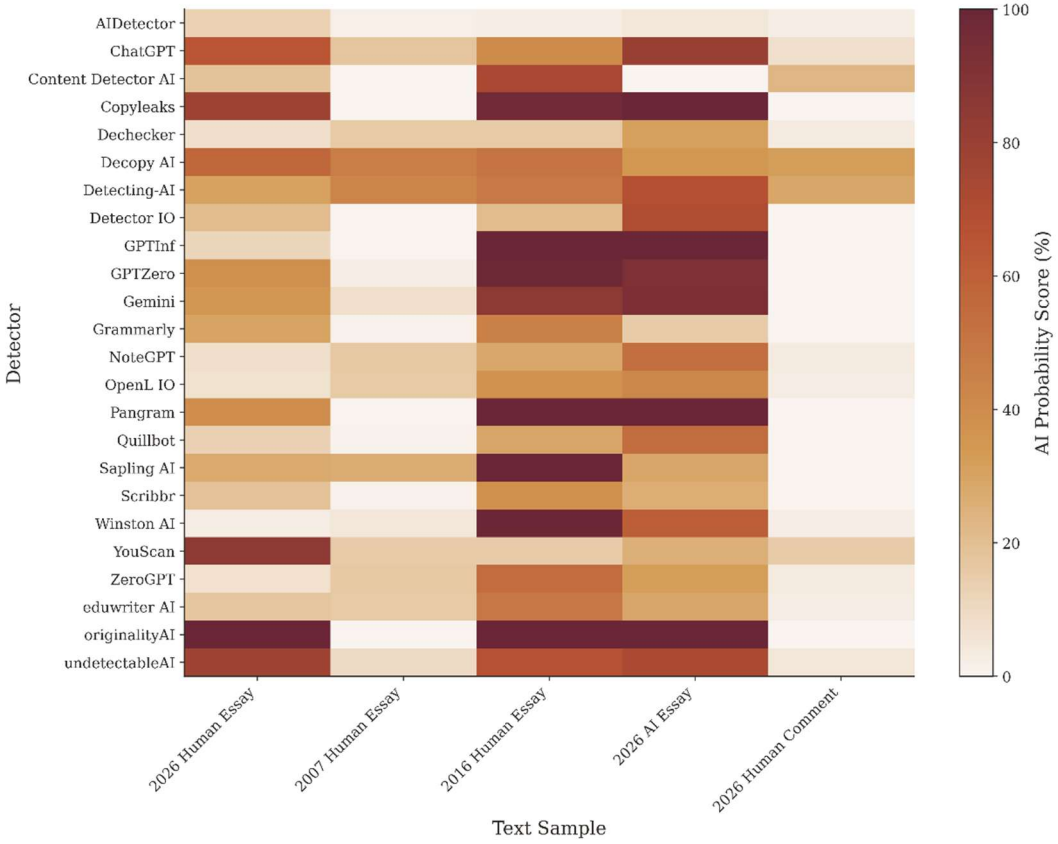


Figure D.1: Detector-by-sample heatmap of raw AI-score outputs (%) across the five texts. With full five-condition profiles now present for every detector, cross-tool disagreement comparable across all conditions without missing-value com

D.2 Detector Performance by Text Sample

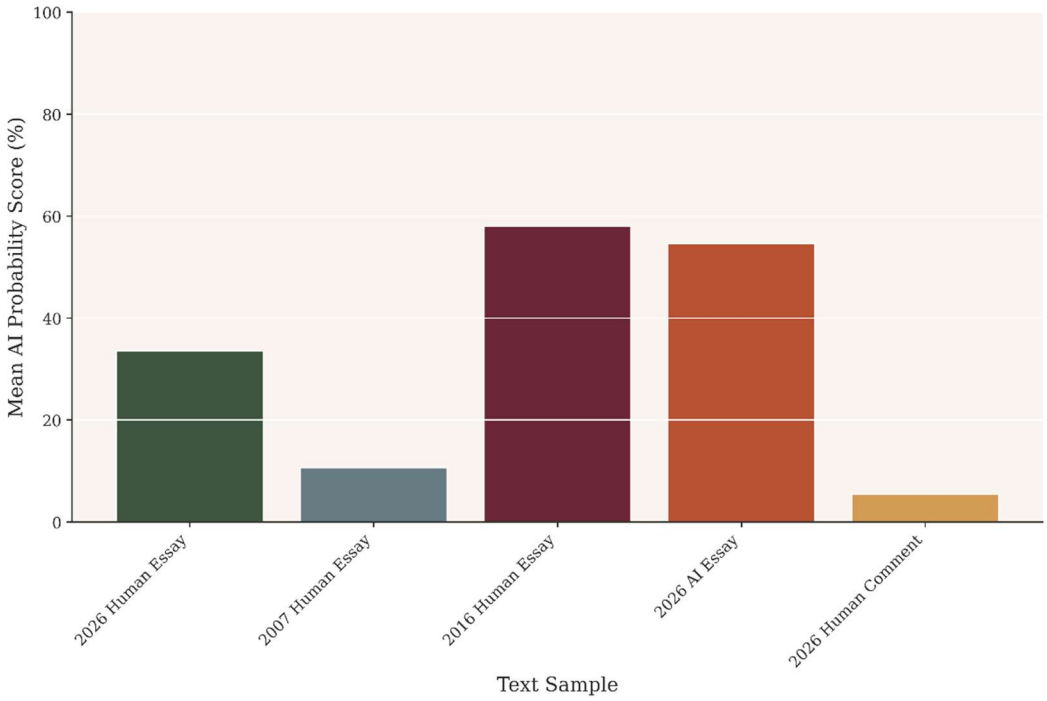


Figure D.2: Mean AI-probability score (%) assigned by detectors to each text sample. Bars represent the arithmetic average of raw AI-likelihood outputs across all detectors for each condition. Values reflect continuous probability scores rather than binary classifications at a fixed decision threshold.

Most detectors gave very low AI scores to the 2007 essay and 2026 comment. Scores sharply increased for the 2016 and 2026 essay, meaning detectors more frequently labeled it as AI even though both were written by a human. Scores declined somewhat for the 2026 essay but remained substantially higher than in 2007. This most likely is due to the non-revised, non-academic narrative style of the 2026 essay.

D.3 Detector Error Scatterplot

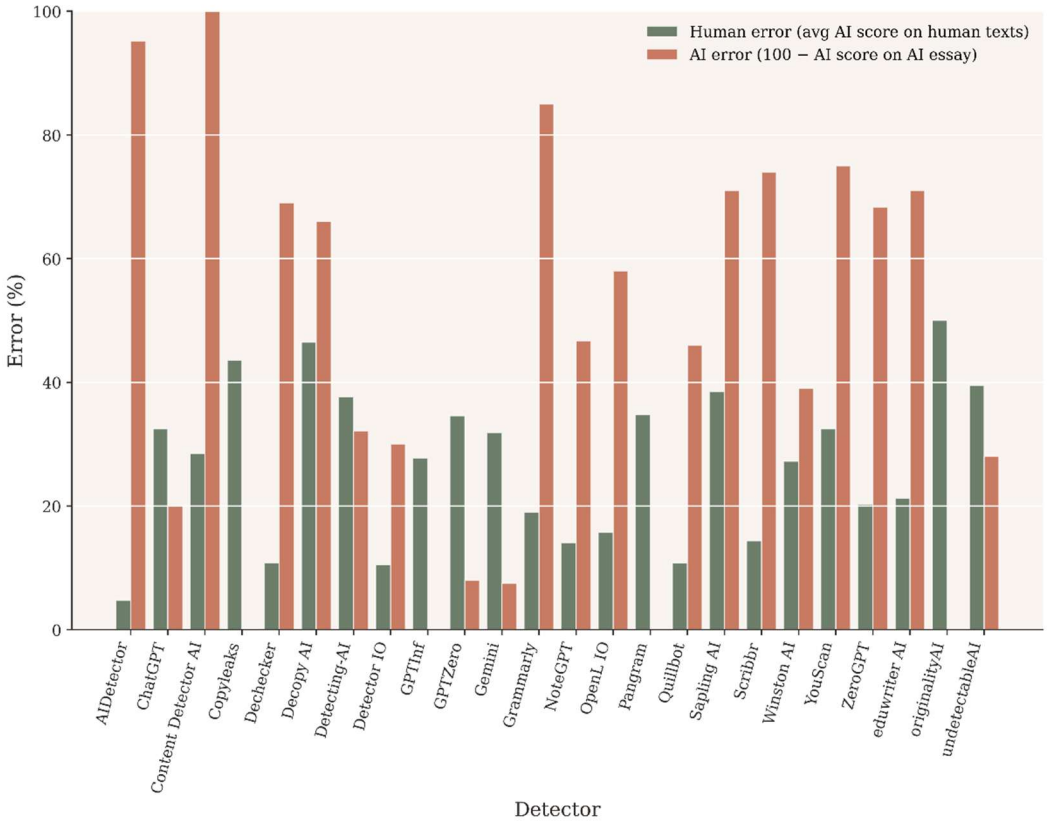


Figure D.3: Per-detector error components. Green bars show the mean AI-probability assigned across all human-written texts (human false-positive tendency). Orange bars show AI detection error, computed as 100 minus the AI-probability assigned to the AI-generated essay (false-negative tendency). Lower combined values indicate more balanced performance across both error types

D.4 Detector Tradeoff Between Accusing Humans and Catching AI.

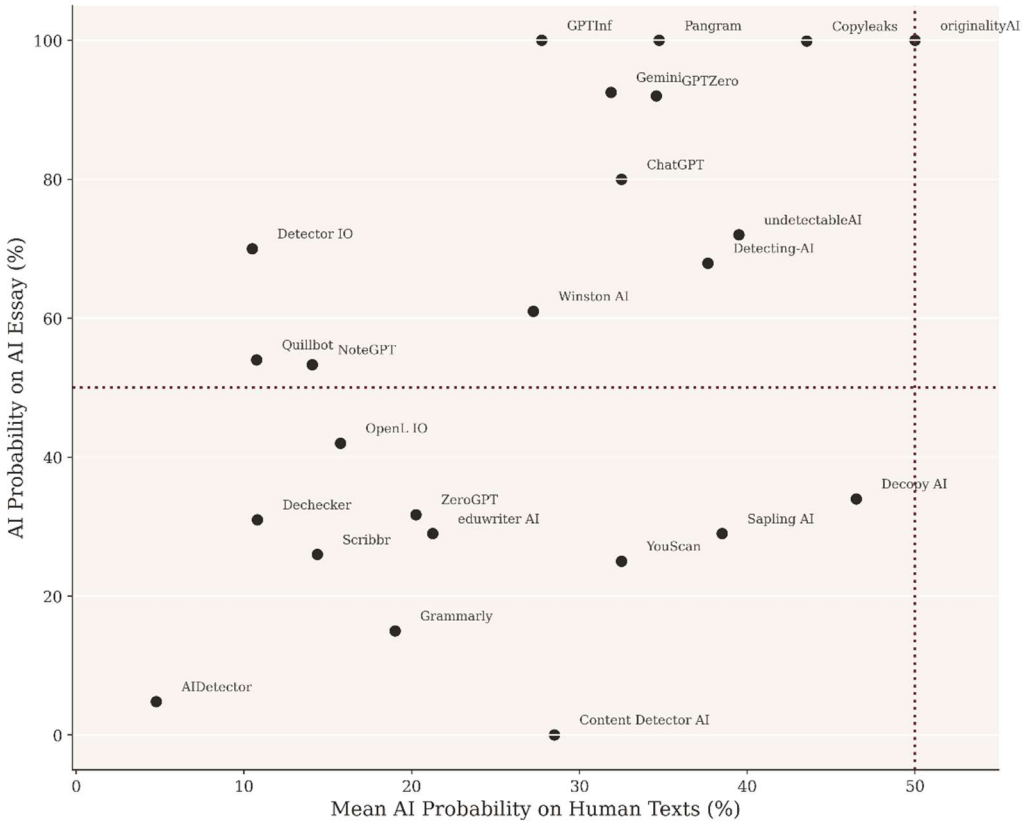


Figure D.4: Each point represents one detector. The horizontal axis shows how strongly it labeled human-written texts as AI. The vertical axis shows how strongly it labeled the actual AI essay as AI. A strong detector would appear in the upper-left area: *low false accusations of humans and high detection of AI*.

Points in other areas show tradeoffs: some tools catch AI but wrongly accuse people, while others avoid accusing people but miss AI. The dashed lines mark the common 50% decision boundary used to classify text as AI.

D.5 Category-Level Error Rates Across Detection Categories

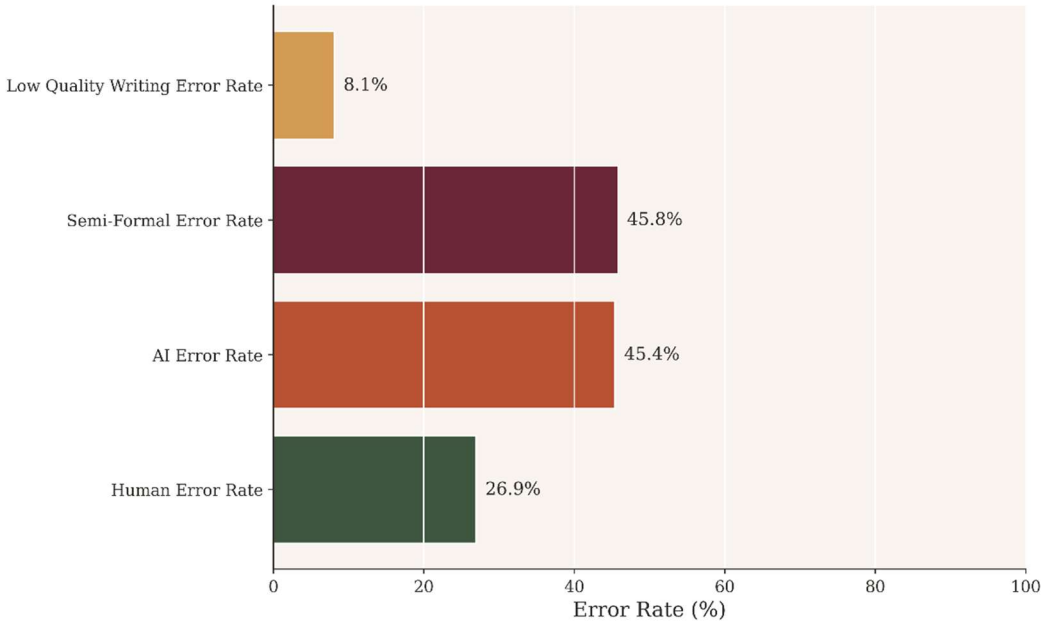


Figure D.5: This figure summarizes four aggregate error categories derived from the detector outputs. The Human Error Rate (26.9%) represents the average probability assigned to AI across all human-written texts. The AI Error Rate (45.4%) reflects the average probability mass not assigned to the AI class when evaluating the AI-generated essay. The Semi-Formal Error Rate (45.8%) captures detector performance on structured human essays, while the Low-Quality Writing Error Rate (8.1%) reflects performance on less structured human writing.